BMC Bioinformatics

RESEARCH Open Access

Structural vs. functional mechanisms of duplicate gene loss following whole genome doubling

David Sankoff[1*], Chunfang Zheng[1], Baoyong Wang[1], Carlos Fernando Buen Abad Najar[2]

From Fourth IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS 2014)
Miami Beach, FL, USA. 2-4 June 2014

Abstract

Background: The loss of duplicate genes - fractionation - after whole genome doubling (WGD) is the subject to a debate as to whether it proceeds gene by gene or through deletion of multi-gene chromosomal segments.

Results: WGD produces two copies of every chromosome, namely two identical copies of a sequence of genes. We assume deletion events excise a geometrically distributed number of consecutive genes with mean $\mu \geq 1$, and these events can combine to produce single-copy runs of length l. If $\mu = 1$, the process is gene-by-gene. If $\mu > 1$, the process at least occasionally excises more than one gene at a time. In the latter case if deletions overlap, the later one simply extends the existing run of single-copy genes. We explore aspects of the predicted distribution of the lengths of single-copy regions analytically, but resort to simulations to show how observing run lengths l allows us to discriminate between the two hypotheses.

Conclusions: Deletion run length distributions can discriminate between gene-by-gene fractionation and deletion of segments of geometrically distributed length, even if μ is only slightly larger than 1, as long as the genome is large enough and fractionation has not proceeded too far towards completion.

Background

The process of whole genome doubling (WGD) gives rise to two copies of each chromosome in a genome, containing the same genes in the same order. Through an attrition mechanism known as fractionation, one of each pair of duplicate genes is lost over evolutionary time. The rare deletion of both copies of a gene can be excluded from our considerations of the interleaving patterns of deletions from duplicated regions first discovered by Wolfe and Shields [1]. The retention of one copy of each pair is what differentiates the WGD/fractionation model from approaches to gene duplication, insertion and deletion in the study of comparative genomics, pioneered by El-Mabrouk [2].

An important biological controversy arises from the question of whether duplicated genes are deleted through random excision - elimination of excess DNA - namely the deletion of chromosomal segments containing one or more genes [3], which we term the "structural" mechanism, or through gene-by gene events such as epigenetic silencing and pseudogenization [4], which are "functional" mechanisms. This question is important to evolutionary theory because it speaks directly to the role of WGD, and gene duplication in general, in disrupting gene order, in creating functional innovation, and in the radiation of new species. It is a question of whether selection operates on the level of simply permitting non-lethal deletions or whether more subtle effects are in play, such as dosage balance of interacting genes.

This debate may be formulated in terms of deletion events removing a number X of contiguous genes, where X is drawn from a geometric distribution γ with mean μ. Here the one-at-a-time deletion model is represented by $\mu = 1$, while the random number of deletions at a time holds if $\mu > 1$. In the latter case, the possibility of two

* Correspondence: sankoff@uottawa.ca
[1]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada
Full list of author information is available at the end of the article

overlapping events is handled by a biologically realistic additive run-length assumption.

In this paper, we investigate the discrimination problem of choosing between the two models based on deletion run-length statistics (resulting from overlapping deletion events). This involves comparing an observed genome containing single-copy genes, originally members of duplicate pairs, to the predictions of the models for $\mu = 1$ and for $\mu > 1$. This requires knowledge of the run-length distribution, given a total number of deleted genes and remaining duplicate pairs. While this is easily calculated for the case $\mu = 1$, the the distribution for the opposing scenario $\mu > 1$ is not known.

In the first part of this paper, we analyze aspects of the deletion run-length distribution ψ when $\mu > 1$ for the deletion-length distribution γ, including some new and surprising analytical results, the clearest of which pertain to a continuous analog of the problem. We then show why it is difficult to describe ψ in closed form or other easily computable format. In the second part, we simulate the distribution and carry out a study of the discrimination problem for various values of μ, genome size N and θ, the proportion of undeleted genes at time t. We conclude with a discussion of the remaining mathematical problems to be solved before the method can be applied to data from real WGD descendants.

Results

The models

For modeling purposes, we consider a doubled genome made up, at the outset, of a pair of identical linear chromosomes each containing genes $g_1, \ldots, g_N$, where N is large enough so that we can neglect end effects - particular behaviors near g_1 *and* g_N . At each time $t = 1, 2, \ldots$, one such doubled gene $\mathrm{g_i}$ is chosen at random, and a value X is chosen from a geometric distribution γ with mean μ. If $X = a$, then $\mathrm{g_i}$, $g_{i+1}, \ldots, g_{i+a-1}$ are deleted from one of the genomes - they become single-copy genes - unless some of these are already single-copy. In the latter case, we skip existing single-copy genes and proceed to convert the next double-copy genes we encounter until a total of a double-copy genes have been converted to single-copy. Note that this overlapping of deletion events never occurs if $\mu = 1$ since, in this case, by definition, exactly one double-copy gene is selected and deleted in each step. For simplicity, we assume all deletions take place from one and the same genome. In a more complete model, deletion events would occur on one or the other chromosome, with probabilities ϕ and $1 - \phi$ [5].

The "skipping" procedure, introduced in [6], is a natural way to model the deletion process, since deletion of part of a chromosome and the subsequent rejoining of the chromosome directly before and directly after the deleted fragment means that this fragment is no longer "visible" to the deletion process. As observers, however, we have a record of the deleted genes, as one copy of each gene must be retained in the genome.

Overlapping deletion events and skipping result in the creation of runs of single-copy genes whose length is the sum of a number of geometric variables. The sum of r identical geometric variables produces a negative binomial distribution with parameter r, but the skipping process does not involve the sum of identical random variables, since a deletion with a large value of a is more likely to overlap an existing single copy region than a deletion with small a. Thus, at any point of time $t > 0$, the distribution ψ_t of single-copy run lengths will tend to contain a higher frequency of runs of length 1, and of very long runs, than would be generated by the negative binomial. On the other hand, the distribution of run lengths of the remaining double-copy genes is geometrically distributed with a probability distribution ρ_t, where the mean ν_t decreases with t [5,6].

Analysis of overlap probabilities

An attempt to determine ψ_t analytically starts with the calculation of how many deletion events have overlapped to form a run of single-copy genes at time t. In [6], we derived a formula to predict whether a deletion event would create a new run of single-copy genes, probability p_0; overlap exactly one existing run, thus extending it without changing the total number of runs, probability p_1; overlap two runs, producing one larger combined run in place of the two pre-existing ones, probability p_2; and so on. Other probabilities deal with the events that a run "touches" a pre-existing run without overlapping it. These probabilities all depend solely on γ and ρ_t. For example, we examine the case of p_0. The other probabilities are all formulated in analogous ways.

The proportion of genes in single-copy runs of length l is $l\rho_t(l)/\nu_t$, where $\nu_t = \sum_{l>0} l\rho_t(l)$. The probability p_0 that a deletion event falls within a run of double-copy genes without deleting the terms at either end is

$$\begin{aligned} p_0 &= \sum_{l>2} \frac{l\rho_t(l)}{\nu_t} \sum_{j=2}^{l-1} \frac{1}{l} \sum_{a=1}^{l-j} \gamma(a) \\ &= \frac{1}{\nu_t} \sum_{l>2} \rho_t(l) \sum_{j=2}^{l-1} \sum_{a=1}^{l-j} \gamma(a) \\ &= \frac{1}{\nu_t} \sum_{l>2} \rho_t(l) \sum_{a=1}^{l-2} (l-a-1)\gamma(a) \end{aligned} \tag{1}$$

where j indexes the starting position of the deletion within a run of length l, and a is the number of genes deleted in the event.

This formula requires quadratic computing time, but the p_i for higher i, require polynomial time of degree $i + 2$. Here we exemplify with p_0 to show that these probabilities can in fact be reduced to closed form, so that computing time is a negligible constant. The formula in (1), when expanded, consists of a number of partial sums of the geometric distributions γ and ρ_t and means of these distributions, all of which are readily reduced to closed form, plus sums of terms of the form $\left[\left(1 - \frac{1}{\mu}\right)\left(1 - \frac{1}{\nu_t}\right)\right]^l$ and $l\left[\left(1 - \frac{1}{\mu}\right)\left(1 - \frac{1}{\nu_t}\right)\right]^l$, which themselves can be considered in terms of a geometric distribution with mean ζ, where

$$1 - \frac{1}{\zeta} = (1 - \frac{1}{\mu})(1 - \frac{1}{\nu_t}) \tag{2}$$

Then (1) reduces to:

$$p_0 = \frac{(\nu_t - 1)^2}{(\mu + \nu_t - 1)\nu_t} \tag{3}$$

For large ν_t, i.e., during the early stage of the process,

$$p_0 \approx \frac{\nu_t}{\mu + \nu_t}\left(1 - \frac{1}{\nu_t}\right) \tag{4}$$

Typically, μ is somewhere between 1 and 2, [3,4], and ν_t of the order of 10^3 or 10^4. Thus p_0 is initially only slightly less than 1 but declines rapidly as ν_t decreases exponentially.

We proceed in an analogous way to derive closed forms for p_1, p_2, . . ., but it is perhaps more instructive here to present the continuous version of the deletion process. Here the two identical chromosomes at time $t = 0$ are linear segments, long enough in comparison with the other parameters of the model so that end effects can be ignored. At each time $t = 1, 2, \ldots$, a random point g is chosen on the chromosome, and a value X is chosen from an exponential distribution

$$f(a) = \frac{1}{\mu} e^{-\frac{a}{\mu}}, a \geq 0 \tag{5}$$

with mean μ. If $X = a$, then the segment $[g, g + a]$ is deleted from one of the genomes - $[g, g + a]$ becomes a single-copy region - unless part of it is already single-copy. In the latter case, we skip existing single-copy regions and proceed to convert the next double-copy region we encounter until a total measure a of double-copy regions have been converted to single-copy.

In analogy with the discrete model, the combined length of the remaining double-copy segments is exponentially distributed according to probability distribution σ_t, with a mean ν_t that decreases with t.

The proportion of undeleted regions accounted for by segments of length ldl is $\frac{l\sigma(l)}{\nu_t}dl$, where $\nu_t = \int_0^\infty l\sigma(l)dl$.

Then the probability p_0 that a deletion event falls completely within an undeleted segment is

$$p_0 = \int_{l=0}^{\infty} \frac{l\sigma_t(l)}{\nu_t} \int_{x=0}^{l} \frac{1}{l} \int_{y=0}^{l-x} f(y)dy\, dx\, dl \tag{6}$$

Carrying out the integrations, we find

$$p_0 = \frac{\nu_t}{\mu + \nu_t} \tag{7}$$

which is reminiscent of the relation (4) in the discrete case with large ν_t.

The probability p_1 that a deletion event overlaps exactly one existing run of deletions is:

$$p_1 = \frac{1}{\nu_t} \int_{l=0}^{\infty} \int_{z=0}^{\infty} \sigma_t(l)\sigma_t(z) \int_{x=0}^{l} \int_{y=l-x}^{l-x+z} f(y)dydxdzdl \tag{8}$$

$$= \frac{\nu_t}{\mu + \nu_t} \frac{\mu}{\mu + \nu_t} \tag{9}$$

It can be proved by induction that the probability a deletion event overlaps exactly q existing runs of deletions is:

$$p_q = \frac{\nu_t}{\mu + \nu_t}\left(\frac{\mu}{\mu + \nu_t}\right)^q \tag{10}$$

Thus we have the surprisingly uncomplicated result that the number q of pre-existing runs of single-copy regions overlapped by a new deletion event is geometrically distributed on q = 0, 1, . . . with parameter $\mu/(\mu + \nu_t)$.

On the run-length distribution

Although having a closed form for p_q constitutes progress towards the computation of the run-length distribution ψ_t, or eventually towards some analytical results on it, how to find this distribution remains a difficult question. As mentioned in the previous section, long deletion events will be involved in more skipping than small ones. This is illustrated in Figure 1, where runs built from a small number of events tend to be composed of shorter deletions, especially when μ is large. Had we just added independent samples from a geometric distribution, the curves in the figure would have been horizontal lines.

How to account for the distorting effect of skipping on the run-length distribution will require additional

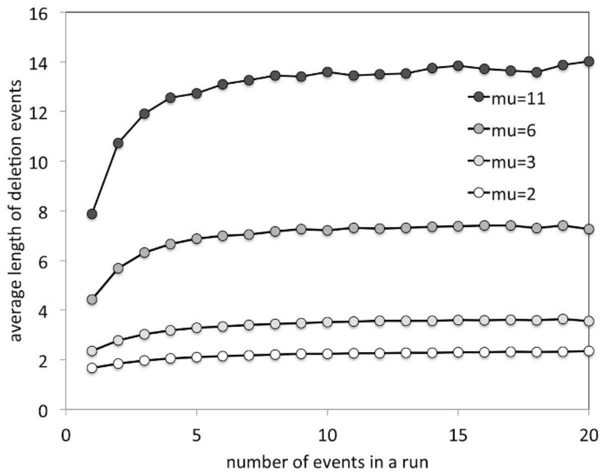

Figure 1 Non-independence of overlapping events. Simulation of the number of overlapping deletion events making up a single-copy region, when 70% of the genes are single copy. With a large number of events in a run, the individual events tend to have greater lengths. From [6].

insight and research. In the interim, we may use simulations to study the discrimination problem.

Simulations

We first simulated the fractionation process for all combinations of the following parameter values:

- gene number N = 100 to 900, in steps of 100.
- μ = 1.0 to 2.4, in steps of 0.1.
- Proportion of the genes deleted, $1 - \theta$ = 0.1 to 0.9, in steps of 0.1.

For each combination of the parameters μ, N and θ, we calculated the distribution of run lengths l for single-copy regions, and similarly for double-copy regions. The simulation was repeated 1000 times and the frequencies of length (l = 1,2,3,...) of runs of deleted genes were averaged over the 1000 trials to get a reasonably accurate estimate of the cumulative $F_{\mu,N,1-\theta}$. Similarly we estimated the cumulative $G_{\mu,N,1-\theta}$ for runs of remaining double-copy genes.

Once the cumulative distributions were established, we then carried out the actual discrimination study. For each value of μ and N, we sampled 1000 new individual trajectories of the deletion process until $1 - \theta$ = 90 % of the genes were deleted. For each value of $1 - \theta$ = 0.1, 0.2, . . . , 0.9, we created "bins" corresponding to the fifteen values of μ for which we had constructed cumulatives. Then for each sample S_i, at each $1 - \theta$ = 0.1, . . . , 0.9 we counted the frequency of runs of deleted genes of length = 1, 2, . . . and constructed a cumulative distribution. We calculated the Kolmogorov-Smirnov statistic $D^{(Si)}_{\mu,N,1-\theta}$ between the sample cumulative and the previously established distribution $F_{\mu,N,1-\theta}$ for each fifteen values of μ and assigned the sample to the bin corresponding to the minimal value of this statistic, which was our estimate $\hat{\mu}$ for that sample.

Figure 2 shows the distributions of $\hat{\mu}$ for the 1000 samples $S_1, \ldots, S_{1000}$. The four panels are the results of N = 900, 300, 200 and 100. A separate distribution is drawn for each of the trial values of μ used to generate the samples. For N = 900 (top left), there is a clear pattern of the mode of the distribution to occur at the same value of μ that generated the data, though the distributions become more spread out for higher values of μ. The same pattern may be seen for N = 300 (top right), though considerably degraded. This loss of accuracy of $\hat{\mu}$ continues through N = 200 (bottom left) and N = 100 (bottom right), where the modes for $\hat{\mu}$ when μ = 1.1 are in the μ = 1.0 bin.

With all four values of N in Figure 2, the most accurate inference is made for μ = 1, the gene-by-gene model. This brings us back to the original problem of discriminating between the gene-by-gene "functional model" (μ = 1) and the random excision "structural" model (μ > 1). Figure 3 shows the frequency with which we estimate $\hat{\mu}$ = 1, for various values of μ and N = 200 or 900, as a function of $1 - \theta$, the proportion of genes deleted. The upper curves in the figure show that we can correctly identify the μ = 1 model around 70-85% of the time; more for N = 900 and less for N = 200, as long as $1 - \theta$ < 50%. In other words, the type I error of a test of $\mathrm{H}_0 : \mu = 1$ against $H_1 : \mu > 1$ with these parameters and procedures, is about 15-30%. The lower curves show that incorrectly inferring $\hat{\mu} = 1$ occurs around 20% of the time when μ = 1.2, but very rarely for μ = 1.9 or even μ = 1.5, until $1 - \theta$ begins to exceed 50%. In other words, if now $H_m : \mu = m$, for some constant $m > 1$, is the null hypothesis and H_0 is the alternative, then the type I error is very small unless m is very close to 1 (e.g., m = 1.2) or $1 - \theta$ is large (e.g., >50% if m = 1.5).

Up to now, we have examined only runs of single-copy genes. What of the runs of remaining double-copy genes? Figure 4 compares some of the results from the same simulations as Figure 3, but using the cumulative $G_{\mu,N,1-\theta}$ for runs of double-copy genes as well as $F_{\mu,N,1-\theta}$ for runs of single-copy genes. The main observation is that the double-copy approach systematically infers μ = 1 with higher frequency for small values of $1 - \theta$, whether or not this inference corresponds to the generating μ. It systematically infers μ = 1 with lower frequency for large values of $1 - \theta$, again whether or not this is correct. These simulations establish ranges of

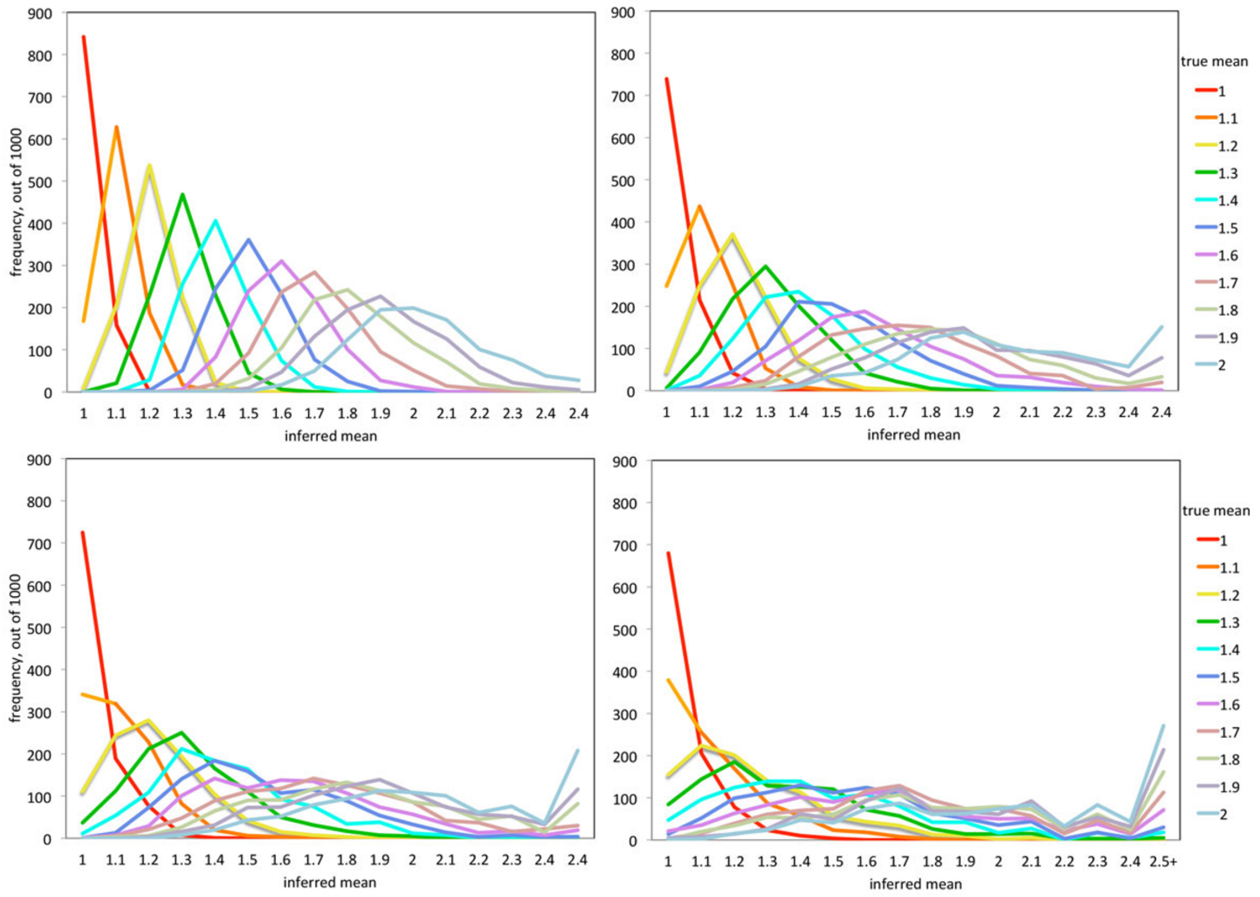

Figure 2 Discrimination between models based on run lengths of deleted genes. Frequency of $\hat{\mu}$, the value for which $D^{(S_i)}_{\mu,N,1-\theta}$ between the sample cumulative and the distribution $F_{\mu,N,1-\theta}$ is minimal. All data involve a proportion of $1 - \theta = 0.20$ deleted genes. Top left: $N = 900$. Top right: $N = 300$. Bottom left: $N = 200$. Bottom right: $N = 100$.

values of N, μ and $1 - \theta$ for which we can and cannot discriminate between the two models.

Conclusions

In this work we have made some progress in deriving the run-length distribution ψ_t for single-copy regions, although this problem is still not completely resolved. From an analytical point of view, it is unexpected and interesting that in the continuous version of the problem, the number of pre-existing runs overlapped by a deletion event follows a geometric distribution.

The simulation study showed the much greater difficulty in distinguishing between the structural and functional models when the mean μ of the deletion size distribution is 1.1 rather than 1.9, when N is 100 rather than 900, and when the proportion of genes deleted is bigger than 50% rather than less than 40%. The latter effect is also apparent in empirical studies [7].

Our simulation results are based on a “binning” strategy for determining $\hat{\mu}$ for the purposes of discrimination, rather than an asymmetrical testing approach comparing the hypotheses $\mu = 1$ and $\mu > 1$. This is justified by the lack of any biological significance, and high rates of error, in comparing $\mu = 1+\in$ and $\mu = 1$ for very small $\in$, as well as the global picture it offers of the degradation of discriminatory power as a function of μ, N and θ.

This work has for the first time enabled the systematic discrimination between the two models of duplicate deletion following WGD. Future research will continue on the analytical determination of ψ_t as well as extension to the “two-sided” deletion models proposed in [5]. Eventually, we will have to allow processes of genome rearrangement to disrupt runs of single-copy genes or double-copy genes, as in [7]. It is these kinds of model that will eventually be useful for analyzing data from real genomes.

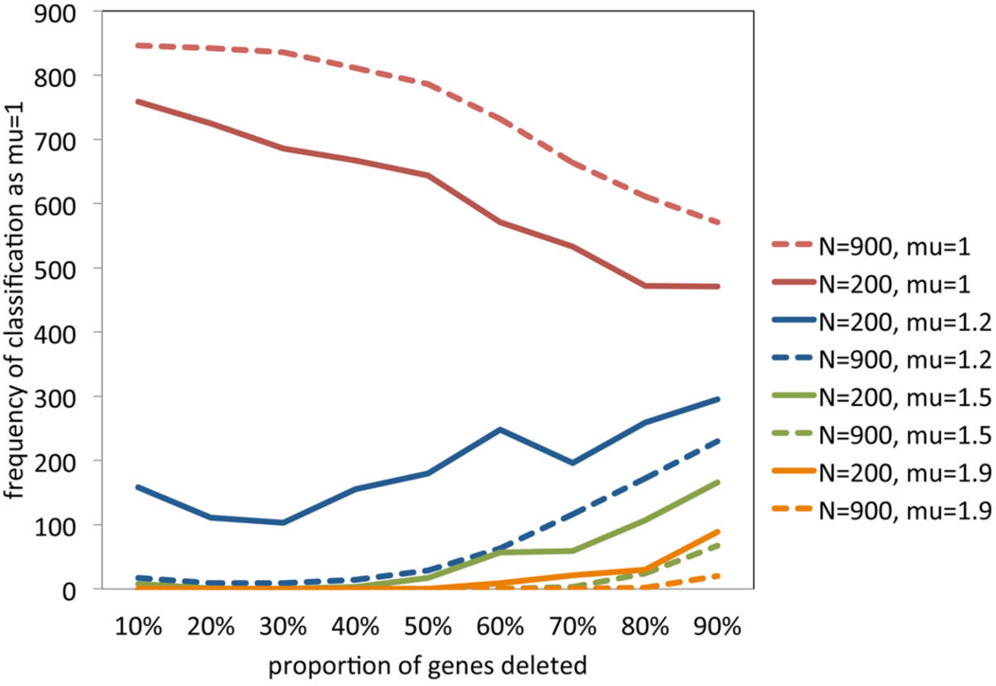

Figure 3 Assessment of tests. Frequency of $\hat{\mu} = 1$ as a function of $1 - \theta$, for $\mu = 1$ (functional hypothesis) and various $\mu > 1$ (structural hypothesis), for $N = 900$ and 200.

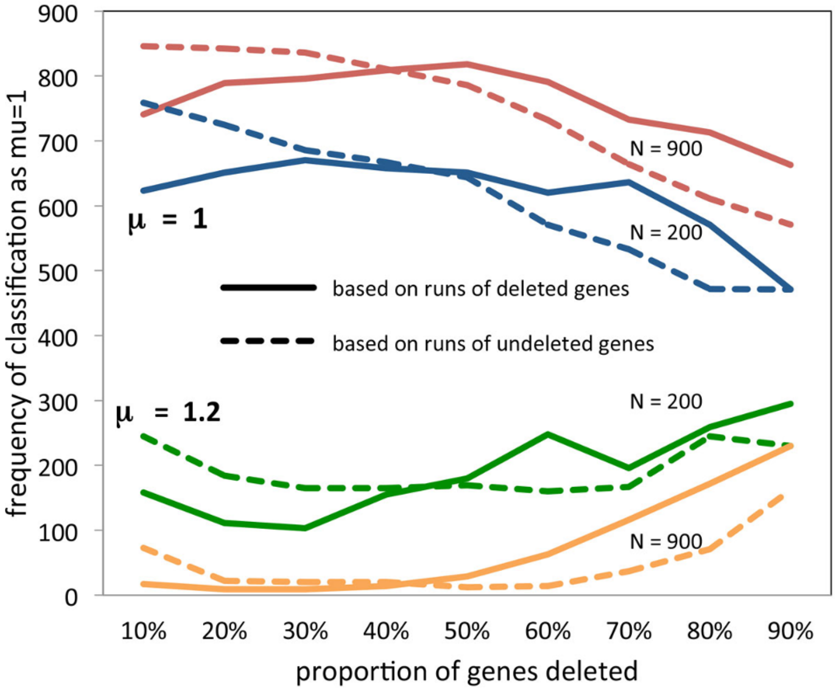

Figure 4 Comparison of double-copy and single-copy analyses. Frequency of $\hat{\mu} = 1$ as a function of $1 - \theta$, for $\mu = 1$ and 1.2 and $N = 900$ and 200. Results based on runs of single-copy (deleted) genes contrasted with results from double-copy (undeleted) genes.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors participated in the research, wrote the paper, read and approved the manuscript.

Acknowledgements

Research supported in part by grants to DS from the Natural Sciences and Engineering Research Council of Canada (NSERC). CFBAN was a Mitacs Globalink intern. DS holds the Canada Research Chair in Mathematical Genomics.

Declarations

Publication of this article was funded by NSERC Discovery grant DG 8867-2008 RGPIN.

This article has been published as part of *BMC Bioinformatics* Volume 16 Supplement 17, 2015: Selected articles from the Fourth IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS 2014): Bioinformatics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/16/S17.

Authors' details

[1]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada. [2]Facultad de Ciencias, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Distrito Federal, México.

Published: 7 December 2015

References

1. Wolfe KH, Shields DC: **Molecular evidence for an ancient duplication of the entire yeast genome.** *Nature* 1997, **387**:708-13.
2. El-Mabrouk N: **Genome rearrangement by reversals and insertions/deletions of contiguous segments.** In *Combinatorial Pattern Matching, 11th Annual Symposium Lecture Notes in Computer Science. Volume 1848.* Springer;Giancarlo, R., Sankoff, D 2000:222-234.
3. van Hoek MJ, Hogeweg P: **The role of mutational dynamics in genome shrinkage.** *Molecular Biology and Evolution* 2007, **24**:2485-2494.
4. Byrnes JK, Morris GP, Li WH: **Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion.** *Molecular Biology and Evolution* 2006, **23**:1136-1143.
5. Sankoff D, Zheng C, Wang B: **A model for biased fractionation after whole genome duplication.** *BMC Genomics* 2012, **13**:S1, S8.
6. Wang B, Zheng C, Sankoff D: **Fractionation statistics.** *BMC Bioinformatics* 2011, **12**:S9, S5.
7. Sankoff D, Zheng C, Zhu Q: **The collapse of gene complement following whole genome duplication.** *BMC Genomics* 2010, **11**:313.

doi:10.1186/1471-2105-16-S17-S9
Cite this article as: Sankoff *et al.*: **Structural vs. functional mechanisms of duplicate gene loss following whole genome doubling.** *BMC Bioinformatics* 2015 **16**(Suppl 17):S9.